All-Pole Modeling of Degraded Speech

JAE S. LIM, STUDENT MEMBER, IEEE, AND ALAN V. OPPENHEIM, FELLOW, IEEE

Abstract—**This paper considers the estimation of speech parameters in an all-pole model when the speech has been degraded by additive background noise. The procedure, based on maximum *a posteriori* (MAP) estimation techniques is first developed in the absence of noise and related to linear prediction analysis of speech. The modification in the presence of background noise is shown to be nonlinear. Two suboptimal procedures are suggested which have linear iterative implementations. A preliminary illustration and discussion based both on a synthetic example and real speech data are given.**

I. INTRODUCTION

PROCESSING of speech which has been degraded due to additive background noise is of interest in a variety of contexts. For example, many speech transmission and coding systems whose design is predicated on a relatively noise-free environment degrade quickly in quality and performance in the presence of background noise [1], [2]. Thus, there is considerable interest in and application for the development of such systems which acknowledge and compensate for the presence of noise.

Furthermore, in many cases, intelligibility is adversely affected by background noise so that a principal objective of a speech processing system may be to improve intelligibility. There have been numerous systems proposed to remove or reduce background noise, with varying degrees of success [3]-[5]. In many cases, these systems provide an apparent improvement in signal-to-noise ratio, but on careful evaluation in fact reduce intelligibility [6]-[8].

Most systems directed at processing speech in the presence of background noise rely, at least to some extent, on a model of the speech waveform as the response of the vocal tract, represented as a quasi-stationary linear system, to a pulse-train excitation for voiced sounds or a noise-like excitation for unvoiced sounds. Thus, for example, for voiced speech the spectrum has a harmonic structure and, if the fundamental frequency is known or can be measured, comb filtering can be applied to remove the noise energy between the harmonics of the speech [3], [4]. While this procedure improves the signal-to-noise ratio, intelligibility tests when the additive noise is white or when it corresponds to a competing speaker have demonstrated that over a wide range of signal-to-noise ratios, intelligibility is in fact reduced by comb filtering [6], [7].

Manuscript received August 30, 1977; revised December 20, 1977. This work was supported by the Advanced Research Projects Agency, monitored by ONR Contract N00014-75-C-0951-NR 049-308.

The authors are with the Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

In this paper we attempt to capitalize more fully on the underlying speech model and develop a complete analysis/synthesis system for which, in the analysis, the synthesizer parameters are estimated from the noisy speech waveform. The basic model used is the representation of the vocal tract as a quasi-stationary all-pole system. This choice is motivated in part by the success of this model in the framework of linear prediction speech analysis for parametric analysis/synthesis of speech in the absence of background noise.

The general class of estimation procedures to be considered is maximum *a posteriori* (MAP) estimation [9]-[11]. As we will see, for speech parameter estimation there are a number of variations of this class of procedures. Furthermore, they are closely related to speech analysis based on linear prediction. In the first part of this paper we review and define these methods for parameter estimation of speech without background noise. With this as a foundation we then consider corresponding estimation procedures with additive background noise. As we will discuss, the estimation procedures which result in linear equations with no background noise become nonlinear when noise is introduced. However, by formulating the problem in an iterative form, two suboptimal systems result which converge and in which the estimation procedure is linear at each iteration.

The overall objectives of this paper are to establish a theoretical foundation for estimation of parameters in an all-pole model in the presence of background noise and to suggest some directions for the simplification of the resulting nonlinear procedures. Future work will include a more complete evaluation of the effect on intelligibility of such systems with natural speech and background noise.

II. STATISTICAL PARAMETER ESTIMATION FOR SPEECH IN THE ABSENCE OF NOISE

Speech can be represented as the response of a linear quasi-stationary system, the vocal tract, to a periodic excitation for voiced speech and a noise-like excitation for unvoiced speech. A commonly used and physically reasonable short-time model for the vocal tract is a linear system for which the transfer function $V(z)$ is all-pole of the form

$$V(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{1}$$

Thus, on a short-time basis, the speech waveform $s(n)$ is assumed to satisfy a difference equation of the form

0096-3518/78/0600-0197$00.75

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + u(n) + e(n) \tag{2}$$

where $u(n)$ is the input excitation to the system and $e(n)$ represents the modeling error in considering the speech generation process as the output of an all-pole system excited by a simplified source. For unvoiced speech, $u(n)$ is random noise. For voiced speech, $u(n)$ over each analysis frame consists of one or several impulses with spacing corresponding to the fundamental pitch period. Throughout our discussion, the basic problem is that of estimating the vocal-tract parameters a_k from a sequence of observations of $s(n)$. Our approach is to consider the combined excitation $u(n) + e(n)$ in (2) as a single noise excitation term which we represent as $g \cdot w(n)$ where g is a gain factor. Strictly speaking, this is only a reasonable representation for the excitation for unvoiced speech. For the case in which there is no background noise our discussion could alternatively be carried out assuming that for voiced speech, $u(n)$ is a known excitation term. If this is carried through, it can be shown, in fact, that because of the specific form of $u(n)$ as one or several impulses, its influence on the estimation procedure is minor. This is substantiated experimentally by virtue of the fact that, as we will see, with the excitation treated as random, one set of estimation procedures corresponds exactly to linear prediction analysis which is well known to be successful for both voiced and unvoiced speech.

Notationally (a summary of notation used throughout the paper is given in the Appendix) it is convenient to represent (2) in matrix form as

$$s(n) = \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p) + g \cdot w(n) \tag{3}$$

where $\boldsymbol{a}$ is the parameter vector

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_p \end{pmatrix} \tag{4}$$

and $\boldsymbol{s}(n_1, n_2)$ denotes the vector of speech samples

$$\boldsymbol{s}(n_1, n_2) = \begin{pmatrix} s(n_1) \\ \cdot \\ \cdot \\ \cdot \\ s(n_2) \end{pmatrix}. \tag{5}$$

Consistent with our discussion above, the excitation $u(n) + e(n)$ has been replaced by $g \cdot w(n)$ where $w(n)$ is taken to be white Gaussian noise with zero mean and unit variance. We assume that the vector of observations consists of N values $s(N-1)$, $s(N-2)$, $\cdots$ $s(0)$, i.e., $\boldsymbol{s}(N-1, 0)$, which we will denote by $\boldsymbol{s}_O$.

From (3) it is clear that $s(n)$ depends on a total of $2p + 1$ parameters, specifically the p values in the coefficient vector $\boldsymbol{a}$, the initial conditions $\boldsymbol{s}(-1, -p)$ which we will denote by $\boldsymbol{s}_I$, and the gain factor g. Our basic approach is to consider all of the unknown parameters as random with associated *a priori* Gaussian probability densities. The class of estimation procedures to be considered are the maximum *a posteriori* (MAP) methods, whereby the basic approach is to choose the parameter estimates to maximize the probability density function of the parameters given the observations [9]-[10].

Given the basic model of (3) and the observation vector $\boldsymbol{s}(N-1, 0)$ we wish to estimate only the coefficient vector $\boldsymbol{a}$. Thus the MAP estimation procedure corresponds to maximizing the probability density function $p(\boldsymbol{a}|\boldsymbol{s}_O)$, which in general, requires the solution of a set of nonlinear equations. Alternatively, we can choose to estimate all of the parameters, i.e., $\boldsymbol{a}$, g, and $\boldsymbol{s}_I$ by maximizing $p(\boldsymbol{a}, g, \boldsymbol{s}_I|\boldsymbol{s}_O)$ even though the only parameters of interest are the coefficients $\boldsymbol{a}$. Thus, several strategies emerge for estimating $\boldsymbol{a}$ depending on how we choose to treat or estimate the remaining parameters. In particular, we consider four cases. In Case 1, we jointly estimate all of the parameters $\boldsymbol{a}$, g and $\boldsymbol{s}_I$ assuming no *a priori* information. The estimate for $\boldsymbol{a}$ that results corresponds exactly to the covariance method of linear prediction analysis. In Case 2 we assume that $\boldsymbol{s}_I$ is known and jointly estimate $\boldsymbol{a}$ and g assuming no *a priori* information. Depending on specifically how we assume $\boldsymbol{s}_I$ is known, this corresponds to estimating $\boldsymbol{a}$ using either the covariance method or the correlation method of linear prediction. In Case 3 we assume that g is known and jointly estimate $\boldsymbol{a}$ and $\boldsymbol{s}_I$, assuming no *a priori* information about $\boldsymbol{s}_I$. In Case 4 we consider estimating only $\boldsymbol{a}$. The resulting set of equations is nonlinear except in the case where $\boldsymbol{s}_I$ and g are both known.

Case 1: In this case, we maximize $p(\boldsymbol{a}, g, \boldsymbol{s}_I|\boldsymbol{s}_O)$ with respect to $\boldsymbol{a}$, g, and $\boldsymbol{s}_I$ with the assumption that no *a priori* information on $\boldsymbol{a}$, g, or $\boldsymbol{s}_I$ is available. This corresponds to the case when $p(\boldsymbol{a}, g, \boldsymbol{s}_I)$ is constant. In obtaining $p(\boldsymbol{a}, g, \boldsymbol{s}_I|\boldsymbol{s}_O)$, we use Bayes' rule to obtain

$$p(\boldsymbol{a}, g, \boldsymbol{s}_I|\boldsymbol{s}_O) = \frac{p(\boldsymbol{s}_O|\boldsymbol{a}, g, \boldsymbol{s}_I)\, p(\boldsymbol{a}, g, \boldsymbol{s}_I)}{p(\boldsymbol{s}_O)}. \tag{6}$$

Since $p(\boldsymbol{s}_O)$ is not a function of $\boldsymbol{a}, g$, or $\boldsymbol{s}_I$ and $p(\boldsymbol{a}, g, \boldsymbol{s}_I)$ is assumed to be constant, maximizing $p(\boldsymbol{a}, g, \boldsymbol{s}_I|\boldsymbol{s}_O)$ in (6) is equivalent to maximizing $p(\boldsymbol{s}_O|\boldsymbol{a}, g, \boldsymbol{s}_I)$.[1] Thus, as is well known [9], MAP estimation of $\boldsymbol{a}$, g, and $\boldsymbol{s}_I$ in the absence of *a priori* information reduces to maximum likelihood (ML) estimation of those parameters. The conditional density function $p(\boldsymbol{s}_O|\boldsymbol{a}, g, \boldsymbol{s}_I)$ can be evaluated by noting that

$$\begin{aligned} p(\boldsymbol{s}_O|\boldsymbol{a}, g, \boldsymbol{s}_I) &= p(\boldsymbol{s}(N-1, 0)|\boldsymbol{a}, g, \boldsymbol{s}(-1, -p)) \\ &= \prod_{n=0}^{N-1} p(s(n)|\boldsymbol{a}, g, \boldsymbol{s}(n-1, -p)) \\ &= \prod_{n=0}^{N-1} p(s(n)|\boldsymbol{a}, g, \boldsymbol{s}(n-1, n-p)). \end{aligned} \tag{7}$$

[1] As the variance becomes larger, the density function becomes wider and flatter approaching a constant. More formally, however, we should assume that $p(\boldsymbol{a}, g, \boldsymbol{s}_I)$ is jointly Gaussian whose covariance approaches an arbitrarily large value. In all cases in this paper where we assume that no *a priori* information of some parameters corresponds to uniform density of the parameters, it can be shown that the same theoretical results are obtained by first solving the case when the variance is finite and then letting the variance approach ∞.

From the model of (3) and the assumption that $w(n)$ is white Gaussian noise with unit variance,

$$p(s(n) \mid \boldsymbol{a}, g, s(n-1), \cdots s(n-p)) = \frac{1}{(2\pi g^2)^{1/2}} \cdot \exp\left[-\frac{1}{2g^2} \cdot (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p))^2\right] \quad (8)$$

and consequently

$$p[\boldsymbol{s}_O \mid \boldsymbol{a}, g, \boldsymbol{s}_I] = \frac{1}{(2\pi g^2)^{N/2}} \cdot \exp\left[-\frac{1}{2g^2} \cdot \sum_{n=0}^{N-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p))^2\right]. \quad (9)$$

Maximizing $p(\boldsymbol{s}_O \mid \boldsymbol{a}, g, \boldsymbol{s}_I)$ with respect to g, we obtain

$$g^2 = \frac{1}{N} \sum_{n=0}^{N-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p))^2. \quad (10)$$

Maximization of $p(\boldsymbol{s}_O \mid \boldsymbol{a}, g, \boldsymbol{s}_I)$ with respect to $\boldsymbol{a}$ and $\boldsymbol{s}_I$ is equivalent to minimizing ϵ_p given by

$$\epsilon_p = \frac{1}{g^2} \sum_{n=0}^{N-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p))^2. \quad (11)$$

Thus we choose the parameters $\boldsymbol{a}$ and $\boldsymbol{s}_I$ to satisfy the set of equations

$$\frac{\partial \epsilon_p}{\partial a_i} = 0 \quad \text{for } i = 1, 2, \cdots, p \quad (12a)$$

$$\frac{\partial \epsilon_p}{\partial s(-j)} = 0 \quad \text{for } j = 1, 2, \cdots, p. \quad (12b)$$

Let us rewrite (11) as

$$\epsilon_p = \frac{1}{g^2} \cdot \sum_{n=0}^{p-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p))^2 + \frac{1}{g^2} \sum_{n=p}^{N-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p))^2. \quad (13)$$

Only the first of these summations involves the initial conditions $\boldsymbol{s}_I$. Furthermore, it is straightforward to show algebraically, and intuitively reasonable, that for any solution of the parameter vector $\boldsymbol{a}$, $\boldsymbol{s}_I$ can be chosen so that the first summation in (13) is zero. Since these are the values which minimize ϵ_p with respect to $\boldsymbol{s}_I$, they would then correspond to our estimate for these parameters. Since we are only interested in explictly estimating the coefficient vector $\boldsymbol{a}$, it is not necessary to solve for $\boldsymbol{s}_I$. Since the first term in (13) will always be zero when ϵ_p is minimized, the minimization of (13) corresponds to minimizing with respect to $\boldsymbol{a}$ the function

$$\frac{1}{g^2} \cdot \sum_{n=p}^{N-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p))^2. \quad (14)$$

Setting the partial derivatives of (14) with respect to each of the coefficients a_i to zero results in a set of linear equations given by

$$\sum_{n=p}^{N-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p)) \cdot s(n-i) = 0, \quad i = 1, \cdots, p. \quad (15)$$

Equation (15) corresponds exactly to the equations obtained by the covariance method of linear prediction analysis [12]-[14].

Case 2: Here, we assume that the initial conditions $\boldsymbol{s}_I$ are known and no *a priori* knowledge of $\boldsymbol{a}$ and g is available. We then maximize $p(\boldsymbol{a}, g \mid \boldsymbol{s}_O; \boldsymbol{s}_I)$ with respect to $\boldsymbol{a}$ and g, where the semicolon separates $\boldsymbol{s}_I$, a known variable, from other variables. Since

$$p(\boldsymbol{a}, g \mid \boldsymbol{s}_O; \boldsymbol{s}_I) = p(\boldsymbol{s}_O \mid \boldsymbol{a}, g; \boldsymbol{s}_I)\, p(\boldsymbol{a}, g; \boldsymbol{s}_I) / p(\boldsymbol{s}_O; \boldsymbol{s}_I) \quad (16)$$

and assuming that $p(\boldsymbol{a}, g; \boldsymbol{s}_I)$ is constant, maximizing $p(\boldsymbol{a}, g \mid \boldsymbol{s}_O; \boldsymbol{s}_I)$ is identical to maximizing $p(\boldsymbol{s}_O \mid \boldsymbol{a}, g; \boldsymbol{s}_I)$ corresponding again to the maximum likelihood estimation of $\boldsymbol{a}$ and g. From (9), maximization of $p(\boldsymbol{s}_O \mid \boldsymbol{a}, g; \boldsymbol{s}_I)$ with respect to g leads to (10) for g^2. Maximization with respect to $\boldsymbol{a}$ is identical to minimizing ϵ_p given by (11). However, the minimization is now carried out with respect to $\boldsymbol{a}$ alone. Comparing (13) and (14), we see that the function to be minimized with respect to $\boldsymbol{a}$ is similar in both cases, differing only in the lower limit of the summation. The linear set of equations for $\boldsymbol{a}$ is now given by

$$\sum_{n=0}^{N-1} (s(n) - \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p)) \cdot s(n-i) = 0, \quad i = 1, \cdots, p. \quad (17)$$

If the initial conditions are indeed known, then we in fact have available $N + p$ observations of $s(n)$, or equivalently we use the first p observations to form the initial condition $\boldsymbol{s}_I$ and the remaining observations to form the observation vector $\boldsymbol{s}_O$. If we consider the relationship between Case 1 and Case 2 on the basis of the same *total* number of observations, then in fact they lead to identical functions to be minimized and consequently identical estimates.

In the above case, we have assumed that $p(\boldsymbol{a}, g; \boldsymbol{s}_I)$ is constant and $\boldsymbol{s}_I$ is exactly known. Therefore, maximization of $p(\boldsymbol{a}, g \mid \boldsymbol{s}_O; \boldsymbol{s}_I)$ was identical to maximizing $p(\boldsymbol{s}_O \mid \boldsymbol{a}, g; \boldsymbol{s}_I)$. Because maximization of $p(\boldsymbol{s}_O \mid \boldsymbol{a}, g; \boldsymbol{s}_I)$ with respect to $\boldsymbol{a}$ and g corresponds to maximum likelihood estimation for $\boldsymbol{a}$ and g given (or conditioned on) the initial conditions $s(-1, -p)$, it is sometimes referred to as the conditional maximum likelihood (CML) estimate of $\boldsymbol{a}$ [14].

As an alternative to using the first p observations in each analysis frame to form the initial condition vector, we can assume that the response was zero prior to the observation interval. In this case, assuming that we have a total of N actual observations, we augment these with p additional zero values. Now, if we further extend the data by p points and augment $s(N + p - 1, N)$ with zeros, then maximization of $p(\boldsymbol{a}, g \mid s(N + p - 1, 0))$ with respect to $\boldsymbol{a}$ and g leads to

$$\sum_{n=0}^{N+p-1} (s(n) - a^T \cdot s(n-1, n-p)) \cdot s(n-i) = 0$$

$$\text{for } i = 1, 2, \cdots, p \tag{18}$$

and $s(N+p-1, N)$ and $s(-1, -p)$ are all 0. This is exactly the same equation given by the correlation method of linear prediction analysis [13]-[16]. In the context of linear prediction analysis, the principal advantage of the correlation method over the covariance method has been that, in that case, the solution of the set of equations involves the inversion of a Toeplitz matrix for which there are particularly efficient methods [17], [18]. In addition, the resulting all-pole model is guaranteed to be stable. From (15) and (18) the resulting linear equations to be solved in both methods are given by

$$\sum_{n} (s(n) - a^T \cdot s(n-1, n-p)) \cdot s(n-i) = 0,$$

$$i = 1, \cdots, p \tag{19}$$

and the summation extends from p to $N-1$ for the covariance method and from 0 to $N+p-1$ for the correlation method.

Case 3: Now we consider the case where g is known so that $p(a, s_I | s_O; g)$ is maximized with respect to a and s_I and no *a priori* information is available about s_I so that $p(s_I | a; g)$ is constant. Again, from Bayes' rule, it follows that

$$\text{maximizing } p(a, s_I | s_O; g) \text{ is equivalent to}$$

$$\text{maximizing } p(s_O | a, g, s_I) \cdot p(a) \tag{20}$$

where we assumed that a is independent of g. Assuming that a has a Gaussian density with mean $\bar{a}$ and covariance function P_0, $p(a)$ is of the form

$$p(a) = \frac{1}{(2\pi)^{p/2} |P_0|^{1/2}} \cdot \exp\left[-\frac{1}{2} (a - \bar{a})^T \cdot P_0^{-1} \cdot (a - \bar{a})\right]. \tag{21}$$

Combining (9), (20), and (21), it can be seen that maximizing (20) is equivalent to minimizing ϵ_p given by

$$\epsilon_p = \frac{1}{g^2} \cdot \sum_{n=0}^{N-1} (s(n) - a^T \cdot s(n-1, n-p))^2 + (a - \bar{a})^T \cdot P_0^{-1} \cdot (a - \bar{a}) \tag{22}$$

ϵ_p in (22) is similar to ϵ_p in (11) or (13) but with the additional term $(a - \bar{a})^T \cdot P_0^{-1} \cdot (a - \bar{a})$. Since this extra term is not a function of s_I, minimization of ϵ_p in (22) with respect to s_I requires that s_I be such that

$$\sum_{n=0}^{p-1} (s(n) - a^T \cdot s(n-1, n-p))^2 = 0.$$

Therefore, minimization of ϵ_p in (22) with respect to a reduces to minimization of ϵ_p given by

$$\frac{1}{g^2} \sum_{n=p}^{N-1} (s(n) - a^T \cdot s(n-1, n-p))^2 + (a - \bar{a}) \cdot P_0^{-1} \cdot (a - \bar{a}). \tag{23}$$

Partial differentiation with respect to a_i for $i = 1, \cdots p$ results in a set of linear equations.

If no *a priori* information on a is assumed so that $P_0 = \sigma^2 I$ with σ^2 arbitrarily large, $\hat{a}$, the estimate of a, obtained in this case is identical to $\hat{a}$ obtained in Case 1.

Case 4: Now we maximize $p(a | s_O)$ with respect to a only. $p(a | s_O)$ can be obtained by integrating out the auxiliary parameters so that

$$p(a | s_O) = \iint_{\text{over } g \text{ and } s_I} p(s_O | a, g, s_I) \cdot \frac{p(a, g, s_I)}{p(s_O)} \cdot dg \cdot ds_I. \tag{24}$$

In general, maximization of $p(a | s_O)$ in (24) leads to a set of nonlinear equations in a. However, it is easy to see that when g and s_I are known, maximizing $p(a | s_O; g, s_I)$ leads to a set of linear equations in a. Furthermore,

$$p(a | s_O; g, s_I) = p(s_O | a; g, s_I) \cdot \frac{p(a; g, s_I)}{p(s_O; g, s_I)} \tag{25}$$

and therefore maximizing $p(a | s_O; g, s_I)$ is equivalent to maximizing $p(s_O | a; g, s_I) \cdot (a; g, s_I)$. Assuming $p(a; g, s_I)$ to be independent of g and s_I and of the form given in (21), maximizing $p(a | s_O; g, s_I)$ in (25) is the same as minimizing the same ϵ_p in (22), which can be easily seen by comparing (20) and (25). Here, however, we minimize ϵ_p with respect to a alone, which again corresponds to solving a set of linear equations. The difference between (22) and (23) is in the limit of the summation, analogous to the difference between (13) and (14). If we assume no *a priori* information of a, then the second term in (23) is eliminated and the estimate for a obtained in this case is identical to that obtained in Case 2.

In the above discussion, we saw that maximizing $p(a | s_O)$ leads to a set of linear equations only when g and s_I are known. In practice we might not expect to know these parameters exactly. However, we might expect to make some reasonable guess of g and s_I. Alternatively, we can solve the linear equations in Case 1; assume that these estimates of g and s_I are exact and maximize (25) with respect to a. A third possibility for obtaining s_I is to use the first p data points as s_I and use the remaining $N-p$ points as s_O, which leads to the same estimate of a as in Case 3.

In our discussion in Case 3 and Case 4, the possibility of incorporating *a priori* information on a was included. Because of the spectral and temporal characteristics of speech, it is possible that *a priori* statistics for a can be developed which in fact aid in the estimation of a. For example, since the vocal tract cannot move arbitrarily fast, the estimate of a in any analysis frame can potentially utilize the result of the estimate in previous frames. This remains an area for study.

For Case 4 for which g and s_I are assumed known it can be shown that $p(a | s_O; g, s_I)$ is Gaussian and thus, in particular, is symmetric about the conditional expectation $E[a | s_O; g, s_I]$. It is well known [11] that in such a case the MAP es-

timate of $\boldsymbol{a}$ is identical to the minimum mean square error estimate. Thus, as an alternative to obtaining the MAP estimate by solving the linear equations obtained by setting the partial derivatives of ϵ_p to zero, a recursive least squares procedure can be used [9], [19], [20]. Such an algorithm corresponds to representing (3) of the speech model by the state equations

$$\boldsymbol{a}(n+1) = \boldsymbol{a}(n)$$
$$s(n+1) = \boldsymbol{s}^T(n, n+1-p) \cdot \boldsymbol{a}(n) + gw(n+1). \tag{26}$$

In (26), $\boldsymbol{a}(n)$ represents the all-pole coefficient vector $\boldsymbol{a}$ at time n. It can be shown [9], [19] that when $\boldsymbol{a}$ is jointly Gaussian, the minimum mean-square error estimate of the states in (26) can be obtained from an iterative solution given by

$$\hat{\boldsymbol{a}}(n+1) = \boldsymbol{k}(n+1) \cdot (s(n+1) - \boldsymbol{s}^T(n, n+1-p) \cdot \hat{\boldsymbol{a}}(n)) + \hat{\boldsymbol{a}}(n) \tag{27}$$

and $\hat{\boldsymbol{a}}(n)$ is the estimate of $\boldsymbol{a}$ based on the *a priori* information of $\boldsymbol{a}$ and the observed data points $s(n, o)$ and $\boldsymbol{k}(n+1)$ is the Kalman filter gain which is a function of g^2, $s(n-1, n-p)$ and the covariance matrix of $\boldsymbol{a}(n)$. The covariance matrix of $\boldsymbol{a}(n)$ can also be updated and the initial starting values $\hat{\boldsymbol{a}}(-1)$ and covariance of $\boldsymbol{a}(-1)$ are, of course, the *a priori* mean and covariance of $\boldsymbol{a}$. For each n, $\hat{\boldsymbol{a}}(n)$ obtained in this manner is identical to $\boldsymbol{a}$ estimated by minimizing the function

$$\frac{1}{g^2} \sum_{l=0}^{n} (s(l) - \boldsymbol{a}^T \cdot s(l-1, l-p))^2 + (\boldsymbol{a} - \bar{\boldsymbol{a}})^T \cdot P_0^{-1} (\boldsymbol{a} - \bar{\boldsymbol{a}}). \tag{28}$$

In particular, $\hat{\boldsymbol{a}}(N-1)$ is the estimate of $\boldsymbol{a}$ obtained from minimizing (22) with respect to $\boldsymbol{a}$.

III. Statistical Parameter Estimation for Speech in the Presence of Noise

In the previous section, we established a framework for MAP parameter estimation of speech in the absence of background noise. In two of its forms, leading to (15) and (18), there has been extensive experience in the context of linear prediction speech analysis with considerable success and they are currently the basis for many speech processing systems [12]-[14], [16]. It is well known, however, that these procedures degrade quickly in the presence of additive background noise [1], [2]. Consequently, it is of interest to consider whether the same basic approach and philosophy can be applied when the observations are recognized to be corrupted by background noise.[2] As we will see, the basic approach, corresponding to Case 4 with the assumption that s_I and g are known, and Cases 1, 2, and 3 in the previous section, which required the solution of a set of linear equations, leads in this case to a set of nonlinear equations, which is generally undesirable. However, as we will discuss, two "suboptimal" procedures can be developed which have linear implementations.

Again, we consider the speech to be generated by the model of (3) and the coefficient vector $\boldsymbol{a}$ as the basic parameters to be estimated. The observation vector $\boldsymbol{y}(N-1, 0)$ consists of the sum of speech and background noise, i.e.,

$$\boldsymbol{y}(N-1, 0) = \boldsymbol{s}(N-1, 0) + \boldsymbol{d}(N-1, 0) \tag{29}$$

where $d(n)$ is background noise. $\boldsymbol{y}(N-1, 0)$ will be alternatively denoted as $\boldsymbol{y}_O$. Further, we assume that $d(n)$ is uncorrelated with $s(n)$ and is generated from a zero-mean white Gaussian process. There is no loss in generality in restricting the additive noise to be white since in the more general case $d(n)$ can be whitened by filtering $y(n)$.

Following a procedure similar to that of Case 4 in the previous section, we can consider choosing the parameters $\boldsymbol{a}$ to maximize $p(\boldsymbol{a} \mid \boldsymbol{y}_O)$. In the previous section when we assumed that g and $\boldsymbol{s}_I$ were known and $p(\boldsymbol{a})$ was Gaussian, the resulting equations were linear. For the current situation, this will no longer be the case. Specifically, from (3) and (29),

$$y(n) = \boldsymbol{a}^T \cdot \boldsymbol{s}(n-1, n-p) + g \cdot w(n) + d(n) \tag{30}$$

or

$$y(n) = \boldsymbol{a}^T \cdot \boldsymbol{y}(n-1, n-p) + g \cdot w(n) + d(n) - \boldsymbol{a}^T \cdot \boldsymbol{d}(n-1, n-p). \tag{31}$$

Expressing $p(\boldsymbol{y}_O \mid \boldsymbol{a}, g, \boldsymbol{s}_I)$ in a manner similar to (7),

$$p(\boldsymbol{y}_O \mid \boldsymbol{a}, g, \boldsymbol{s}_I) = \prod_{n=p}^{N-1} p(y(n) \mid \boldsymbol{a} \cdot g \cdot \boldsymbol{s}_I, \boldsymbol{y}(n-1, 0)) \cdot \prod_{n=1}^{p-1} p(y(n) \mid \boldsymbol{a}, g, \boldsymbol{s}_I, \boldsymbol{y}(n-1, 0)) \cdot p(y(0) \mid \boldsymbol{a}, g, \boldsymbol{s}_I). \tag{32}$$

From (31), for $n \geq p$, $p(y(n) \mid \boldsymbol{a}, g, \boldsymbol{s}_I, \boldsymbol{y}(n-1, 0))$ is Gaussian with mean of $\boldsymbol{a}^T \cdot \boldsymbol{y}(n-1, n-p) - \boldsymbol{a}^T \cdot E[\boldsymbol{d}(n-1, n-p) \mid \boldsymbol{a}, g, \boldsymbol{s}_I, \boldsymbol{y}(n-1, 0)]$ and variance of $g^2 + \sigma_d^2 + \boldsymbol{a}^T \cdot \mathrm{Var}[\boldsymbol{d}(n-1, n-p) \mid \boldsymbol{a}, g, \boldsymbol{s}_I, \boldsymbol{y}(n-1, 0)] \cdot \boldsymbol{a}$ where $E[\boldsymbol{d}(n-1, n-p) \mid \boldsymbol{a}, g, \boldsymbol{s}_I, \boldsymbol{y}(n-1, 0)]$ and $\mathrm{Var}[\boldsymbol{d}(n-1, n-p) \mid \boldsymbol{a}, g, \boldsymbol{s}_I, \boldsymbol{y}(n-1, 0)]$ denote the mean and variance of $\boldsymbol{d}(n-1, n-\mathrm{p})$ conditioned on $\boldsymbol{a}, g, \boldsymbol{s}_I$ and $\boldsymbol{y}(n-1, 0)$. Since the variance is a function of $\boldsymbol{a}$, and will likewise be so for the remaining terms, the resulting equations for maximizing $p(\boldsymbol{a} \mid \boldsymbol{y}_O)$ will by necessity be nonlinear.

Even though we have only shown that maximizing $p(\boldsymbol{a} \mid \boldsymbol{y}_O)$ which corresponds to Case 4 is a nonlinear problem, it is easy to see that maximizing $p(\boldsymbol{a}, g, \boldsymbol{s}_I \mid \boldsymbol{y}_O)$, $p(\boldsymbol{a}, g \mid \boldsymbol{y}_O)$ or $p(\boldsymbol{a}, \boldsymbol{s}_I \mid \boldsymbol{y}_O)$ corresponding to Cases 1, 2, and 3 in the previous section is also a nonlinear problem. This is partly because each of the three density functions $p(\boldsymbol{a}, g, \boldsymbol{s}_I \mid \boldsymbol{y}_O)$, $p(\boldsymbol{a}, g \mid \boldsymbol{y}_O)$, or $p(\boldsymbol{a}, \boldsymbol{s}_I \mid \boldsymbol{y}_O)$ is a product of several terms, one of which is

[2] There are many techniques which have been explored for modeling a system on the basis of noisy observations. Many of these techniques [22], [23] differ in a variety of ways including choice of error criterion, assumed form for the model, assumed knowledge about the input, etc. We restrict our discussion to an all-pole model driven by white noise.

$$\prod_{n=p}^{N-1} p(y(n) \mid a, g, s_I, y(n-1, 0)).$$

It was shown above that

$$p(y(n) \mid a, g, s_I, y(n-1, 0))$$

for $p \leqslant n \leqslant N-1$ has variance which is a function of a.

As an alternative to true MAP estimation of a which is obtained by maximizing $p(a \mid y_O)$, one can consider a "suboptimal" procedure which is computationally more tractable. Specifically, we know from Section II, Case 4 how to estimate a from s_O from a linear set of equations assuming s_I and g are known. In the case of speech without background noise, when s_I is assumed known it is either extracted from the observations or (artificially) taken to be zero. In the current situation the first option is no longer available. As we will see shortly, however, by initially assuming that s_I is known, an approximate procedure develops in which s_I does not play an essential role.

Although we, of course, do not have s_O, we can likewise consider MAP estimation of s_O from the observations y_O, given the coefficients a. This then suggests an iterative procedure whereby we begin with an assumed set of initial values a_0 for the coefficient vector a and based on this, estimate s_O by maximizing $p(s_O \mid a_0, y_O; g, s_I)$. Denoting this first estimate of s_O by $\hat{s}_{O1}$, we then form a first estimate $\hat{a}_1$ of a. This procedure can then be continued iteratively to obtain the final estimate $\hat{a}_\infty$ of the coefficients. It is straightforward to see that this procedure for estimating a (and s_O) converges to a local maximum of the joint probability density $p(a, s_O \mid y_O; g, s_I)$. Specifically, since $\hat{a}_i$ is obtained by maximizing $p(a \mid \hat{s}_{Oi}, y_O; g, s_I)$,

$$p(a \mid \hat{s}_{Oi}, y_O; g, s_I) \mid_{a=\hat{a}_i} \cdot p(s_O \mid y_O; g, s_I) \mid_{s_O=\hat{s}_{Oi}} \geqslant p(a \mid \hat{s}_{Oi}, y_O; g, s_I) \mid_{a=\hat{a}_{i-1}} \cdot p(s_O \mid y_O; g, s_I) \mid_{s_O=\hat{s}_{Oi}}$$

so that

$$p(a, s_O \mid y_O; g, s_I) \mid_{a=\hat{a}_i,\, s_O=\hat{s}_{Oi}} \geqslant p(a, s_O \mid y_O; g, s_I) \mid_{a=\hat{a}_{i-1},\, s_O=\hat{s}_{Oi}} . \tag{33}$$

Furthermore,

$$p(s_O \mid \hat{a}_{i-1}, y_O; g, s_I) \mid_{s_O=\hat{s}_{Oi}} \cdot p(a \mid y_O; g, s_I) \mid_{a=\hat{a}_{i-1}} \geqslant p(s_O \mid \hat{a}_{i-1}, y_O; g, s_I) \mid_{s_O=\hat{s}_{Oi}} \cdot p(a \mid y_O; g, s_I) \mid_{a=\hat{a}_{i-1}}$$

so that

$$p(a, s_O \mid y_O; g, s_I) \mid_{a=\hat{a}_{i-1},\, s_O=\hat{s}_{Oi}} \geqslant p(a, s_O \mid y_O; g, s_I) \mid_{a=\hat{a}_{i-1},\, s_O=\hat{s}_{Oi-1}} . \tag{34}$$

From (33) and (34)

$$p(a, s_O \mid y_O; g, s_I) \mid_{a=\hat{a}_i,\, s_O=\hat{s}_{Oi}} \geqslant p(a, s_O \mid y_O; g, s_I) \mid_{a=\hat{a}_{i-1},\, s_O=\hat{s}_{Oi-1}} . \tag{35}$$

If the initial guess for a is such that the local maximum to which this procedure converges is in fact the global maximum, which will always be the case if $p(a, s_O \mid y_O; g, s_I)$ is unimodal, then this procedure will in fact correspond to that joint MAP estimate of the parameters a and s_O. Thus, in essence, this attempt to simplify the problem computationally corresponds to augmenting the desired set of parameters a with the additional parameters s_O which are really unwanted parameters in the sense that we are not particularly interested in explictly estimating them.

From our discussion in Section II, we know that maximizing $p(a \mid \hat{s}_{Oi}, y_O; g, s_I)$ requires the solution of a set of p linear equations for a. Maximizing $p(s_O \mid \hat{a}_i, y_O; g, s_I)$ requires the solution of a set of N linear equations corresponding to the N values in the vector s_O. Specifically, from Bayes' rule, $p(s_O \mid \hat{a}_i, y_O; g, s_I)$ can be denoted as

$$p(s_O \mid \hat{a}_i, y_O; g, s_I) = p(y_O \mid \hat{a}_i, s_O; g, s_I) \cdot \frac{p(s_O \mid \hat{a}_i; g, s_I)}{p(y_O \mid \hat{a}_i; g, s_I)} . \tag{36}$$

Denoting

$$p(y_O \mid \hat{a}_i, s_O; g, s_I) = \prod_{n=1}^{N-1} p(y(n) \mid \hat{a}_i, s_O, y(n-1, 0); g, s_I) \cdot p(y(0) \mid \hat{a}_i, s_O; g, s_I)$$

and noting that $p(y(n) \mid \hat{a}_i, s_O, y(n-1, 0); g, s_I)$ is Gaussian with mean of $s(n)$ and variance of σ_d^2 for $1 \leqslant n \leqslant N-1$ and $p(y(0) \mid \hat{a}_i, s_O; g, s_I)$ is Gaussian with mean of $s(0)$ and variance of σ_d^2, $p(y_O \mid \hat{a}_i, s_O; g, s_I)$ can be denoted as

$$p(y_O \mid \hat{a}_i, s_O; g, s_I) = \frac{1}{(2\pi\sigma_d^2)^{N/2}} \cdot \exp\left(-\frac{1}{2\sigma_d^2} \cdot \sum_{n=0}^{N-1} (y(n) - s(n))^2\right). \tag{37}$$

Combining (9) and (37) with (36) and noting that $p(y_O \mid \hat{a}_i; g, s_I)$ is not a function of s_O,

$$p(s_O \mid \hat{a}_i, y_O; g, s_I) = \text{constant} \cdot \frac{1}{(4\pi^2 g^2 \cdot \sigma_d^2)^{N/2}} \cdot \exp\left(-\frac{1}{2}\epsilon_p\right) \tag{38a}$$

and

$$\epsilon_p = \frac{1}{g^2} \cdot \sum_{n=0}^{N-1} (s(n) - \hat{a}_i^T \cdot s(n-1, n-p))^2 + \frac{1}{\sigma_d^2} \cdot \sum_{n=0}^{N-1} (y(n) - s(n))^2 . \tag{38b}$$

Maximizing $p(s_O \mid \hat{a}_i, y_O; g, s_I)$ is equivalent to minimizing ϵ_p in (38b), and thus we choose s_O that satisfy the set of linear equations

$$\frac{\partial \epsilon_p}{\partial s(i)} = 0 \quad \text{for } i = 0, 1, \cdots, N-1. \tag{39}$$

In the above, we have reduced the nonlinear problem to the successive solution of sets of linear equations. N, however, will generally be large ($>>p$), perhaps on the order of several hundred. Consequently, the problem is still computationally tedious. However, since ϵ_p in (38b) can be written as

$$\epsilon_p = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \beta_{ij} (s(i) - m_i) \cdot (s(j) - m_j) + \text{constant} \tag{40}$$

and $[\beta_{ij}]^{-1}$ is a covariance matrix, $p(s_O \mid \hat{a}_i, y_O)$ is jointly Gaussian in s_O [21], and consequently the MAP estimate of s_O, based on maximizing this probability density, is equivalent to the MMSE estimate of s_O. Furthermore, as N increases, the procedure for obtaining the MMSE estimate of $s(n)$ given by (39) approaches a noncausal Wiener filter, i.e., $s(n)$ is estimated by filtering $y(n)$ with a linear, time-invariant filter with frequency response

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + \sigma_d^2} \tag{41a}$$

where $P_s(\omega)$ is the power density spectrum of $s(n)$ given $\hat{a}_i$ and g, or

$$P_s(\omega) = \frac{g^2}{\left|1 - \sum_{k=1}^{p} a_k \cdot e^{-jk\omega}\right|^2} \tag{41b}$$

and $\boldsymbol{a}$ in (41b) corresponds to $\hat{a}_i$. Thus, we can approximate the MAP estimate $\hat{s}_{Oi}$ by applying the filter of (41) to the set of observations y_O, using the values $\hat{a}_{i-1}$ for the coefficient vector to determine $P_s(\omega)$. To obtain the estimate $\hat{a}_i$ from $\hat{s}_{Oi}$ we can either use the first p values as the initial condition vector, or always assume that $s_I = 0$.

Based on the same philosophy, there are other approximate systems that appear to be plausible. For example, from Case 4 in Section II we note that the MAP estimate of $\boldsymbol{a}$ corresponding to maximizing $p(\boldsymbol{a} \mid s_O; g, s_I)$ uses the values s_O to form products of the form $s(i) \cdot s(j)$. The use of (41) to estimate s_O corresponds to estimating $s(i) \cdot s(j)$ as

$$\hat{s(i) \cdot s(j)} = E[s(i) \mid \hat{a}_i, y_O; g, s_I] \cdot E[s(j) \mid \hat{a}_i, y_O; g, s_I] \tag{42}$$

since, as is well known, the MMSE estimate of a vector s_O is given by the conditional expectation. As an alternative, we can consider generating directly the MMSE estimate of the product $s(i) \cdot s(j)$. In this case, then, the estimate of $s(i) \cdot s(j)$ is given by

$$\hat{s(i) \cdot s(j)} = E[s(i) \cdot s(j) \mid \hat{a}_i, y_O; g, s_I]. \tag{43}$$

From (40), the covariance matrix of s_O conditioned on $\boldsymbol{a}$ and y_O is given by $[\beta_{ij}]^{-1}$, where the $N \times N$ matrix $[\beta_{ij}]$ is given by (38b) and (40), and $[\beta_{ij}]^{-1}$ represents the inverse of the matrix $[\beta_{ij}]$. Denoting this covariance matrix as $[\gamma_{ij}]$

$$E[s(i) \cdot s(j) \mid \hat{a}_i, y_O; g, s_I] = \gamma_{ij} + E[s(i) \mid \hat{a}_i, y_O; g, s_I] \cdot E[s(j) \mid \hat{a}_i, y_O; g, s_I]. \tag{44}$$

Equation (44) is simple to evaluate once $[\gamma_{ij}]$ is computed. Even though the computation of $[\gamma_{ij}]$ generally requires the inversion of an $N \times N$ matrix, a computationally simpler procedure can be obtained as N becomes large. In particular, as N increases, if we assume γ_{ij} depends only on the time difference $i - j$, we can denote γ_{ij} by $\gamma(n) = \gamma(i - j)$. It can then be shown that $\Gamma(\omega)$, the Fourier transform of $\gamma(n)$, is given by

$$\Gamma(\omega) = \frac{P_s(\omega) \cdot \sigma_d^2}{P_s(\omega) + \sigma_d^2} \tag{45}$$

and $P_s(\omega)$ is given by (41b). Compared with the first method discussed, in which s_O is estimated by means of a Wiener filter, it is not known theoretically whether this method converges. However, as will be discussed in the next section, convergence has been empirically observed in the examples considered.

IV. Preliminary Results

A careful study of the two "suboptimal" linear implementations discussed in the previous section requires an evaluation of its effect on intelligibility and quality when applied to speech degraded by background noise, and such a study is currently being carried out. As a very preliminary illustration of the two methods, they were applied to synthetic data and real speech data with additive background noise. Specifically, the synthetic data are based on a sampling rate of 10 kHz and were generated by exciting a tenth-order all-pole filter whose coefficients were derived from unvoiced segment of real speech. The excitation was chosen in one set of examples to be white noise and in the other set of examples to be a periodic impulse train. As we had discussed previously, the results in this paper are derived assuming a stochastic excitation. For speech without background noise, systems derived from this point of view perform well even when the excitation is a periodic impulse train, and it is anticipated that this will also be true for the systems discussed in this paper. The real speech data used in the example of this paper are the sentence "line up at the screen door" spoken by an adult male speaker. The real speech data were low-pass filtered at 4.8 kHz and sampled at 10 kHz.

The synthetic data are analyzed for the case of zero-mean white Gaussian background noise at three different S/N ratios; 20, 10, and 0 dB. The real speech data are analyzed for the same background noise at S/N ratio of 10 dB. The S/N ratio is defined as $10 \log (\Sigma_n s^2(n)/\Sigma_n d^2(n))$ where, for the case of the synthetic data, the summation is over the length of the analysis segment for which we used the length of 25.6 ms (256 data points). For the real speech data, the summation is over the entire length of the sentence. Three systems are considered.

System A: This system corresponds to the assumption of no background noise and maximizes $p(\boldsymbol{a} \mid s(N + p - 1, 0); g)$ with the assumption that no *a priori* information for $\boldsymbol{a}$ is available and $s(-1, -p)$ and $s(N + p - 1, N)$ are $\mathbf{0}$. Thus, it corresponds to the correlation method of (19).

System B: In this system, the iterative procedure corre-

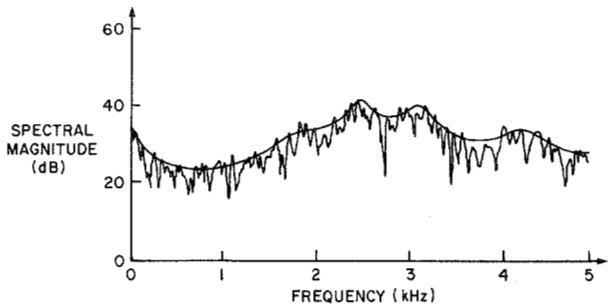

Fig. 1. True spectrum with white noise excitation and vocal tract transfer function estimated by System A.

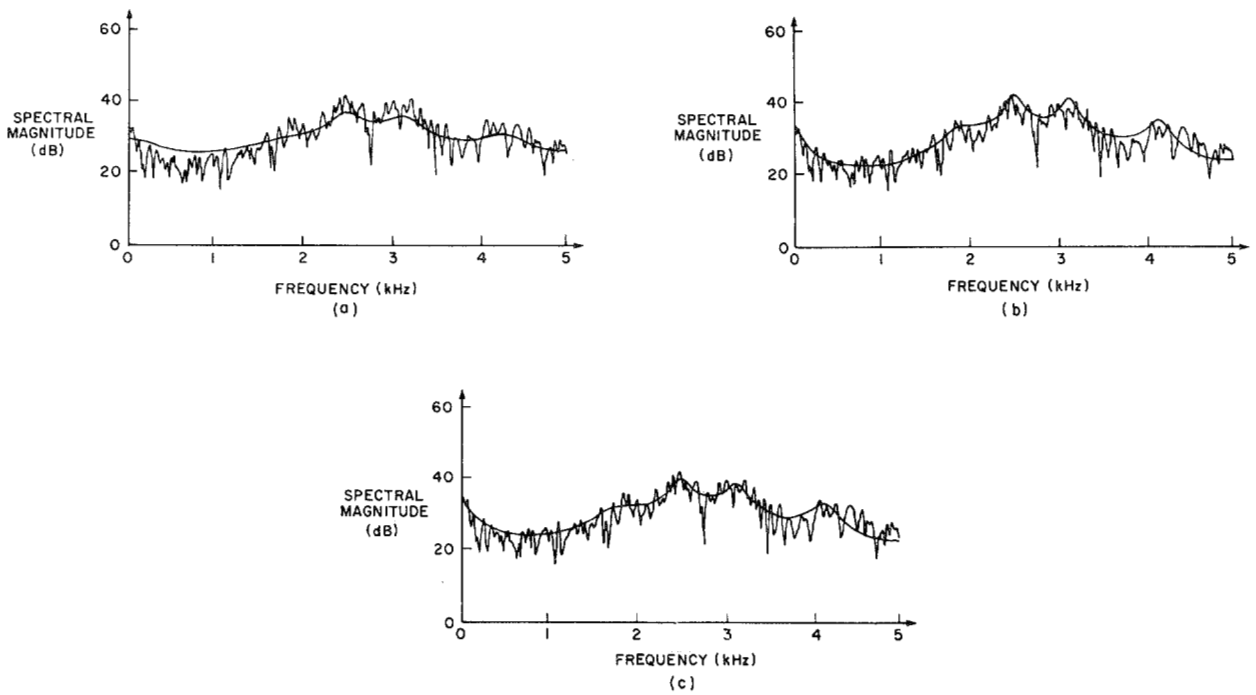

Fig. 2. (a) True spectrum with white noise excitation and vocal tract transfer function estimated by applying System A to noisy synthetic data at S/N of 20 dB. (b) True spectrum with white noise excitation and vocal tract transfer function estimated by applying System B to noisy synthetic data at S/N of 20 dB. (c) True spectrum with white noise excitation and vocal tract transfer function estimated by applying System C to noisy synthetic data at S/N of 20 dB.

sponding to the use of the Wiener filter of (41) to estimate s_O and then System A above to estimate $\boldsymbol{a}$ are applied on each iteration. The initial estimate $\boldsymbol{a}_0$ of the coefficient vector is taken as the result of applying System A to noisy speech. In using (41) to estimate s_O from $\hat{\boldsymbol{a}}_{i-1}$, g was estimated by an energy measurement. More specifically, in all cases, including the case when the excitation is assumed to be a periodic train of impulses, g was obtained from

$$\frac{N}{2\pi}\int_{-\pi}^{\pi} \frac{g^2}{\left|1 - \sum_{k=1}^{p} a_k \cdot e^{-jk\omega}\right|^2} \cdot d\omega$$

$$= \sum_{n=0}^{N-1} y^2(n) - N \cdot \sigma_d^2$$

where $\boldsymbol{a}$ in this equation corresponds to $\hat{\boldsymbol{a}}_{i-1}$ and N is the length of analysis segment or 256 in the examples.

System C: This system is identical to System B except that when $s(i) \hat{\cdot} s(j)$ is computed on each iteration of System B, γ_{ij} obtained by (45) is included to satisfy (44). The initial estimate of the coefficient vector $\boldsymbol{a}_0$ and g was obtained in the same manner as in System B.

Figs. 1-8 show the results for each analysis with the two different forms of excitations and three different S/N ratios. Figs. 1-4 correspond to the case when the excitation is white

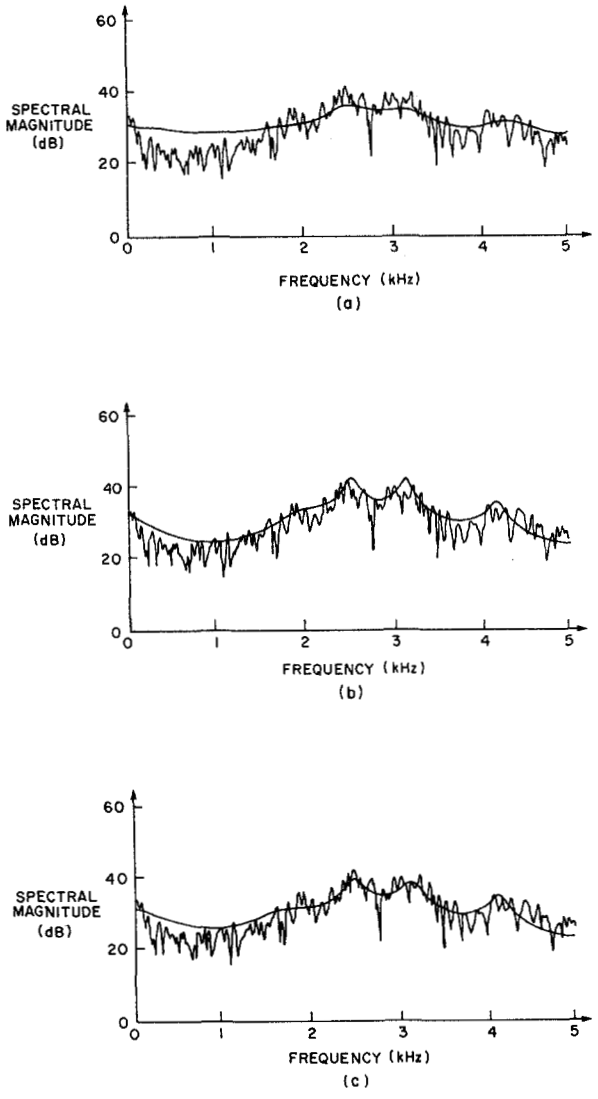

Fig. 3. Same as Fig. 2 with S/N of 10 dB.

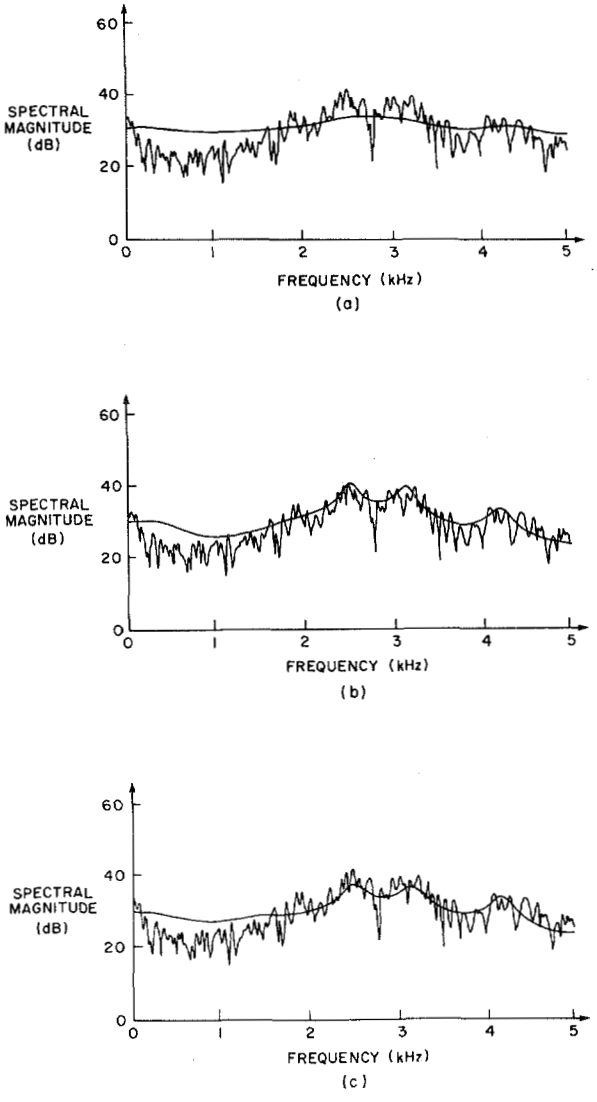

Fig. 4. Same as Fig. 2 with S/N of 0 dB.

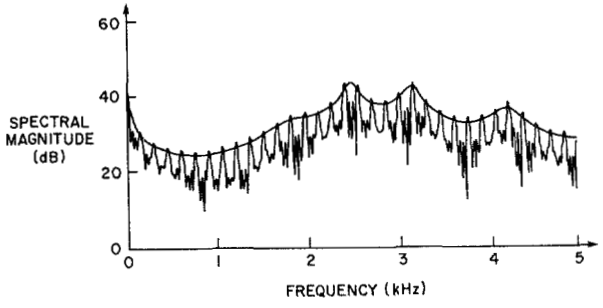

Fig. 5. Same as Fig. 1 with periodic impulse train excitation.

noise and Figs. 5–8 correspond to the case when the excitation is a periodic train of impulses. In Fig. 1, the true spectrum corresponding to white noise excitation and a tenth-order all-pole fit to the spectrum by System A is shown. Figs. 2–4 correspond to S/N ratios of 20, 10, and 0 dB, respectively. In each of the three figures, (a), (b), and (c) represent the estimated vocal-tract transfer function obtained by applying System A to noisy synthetic data (with the

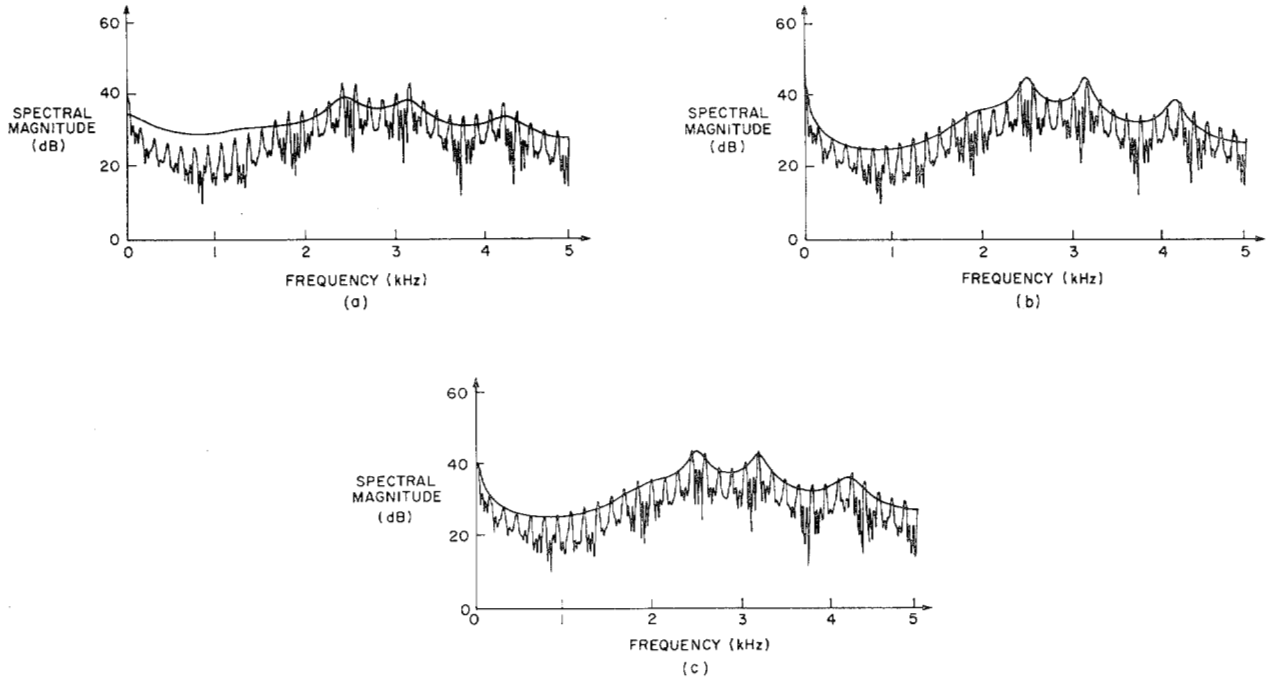

Fig. 6. Same as Fig. 2 with periodic impulse train excitation.

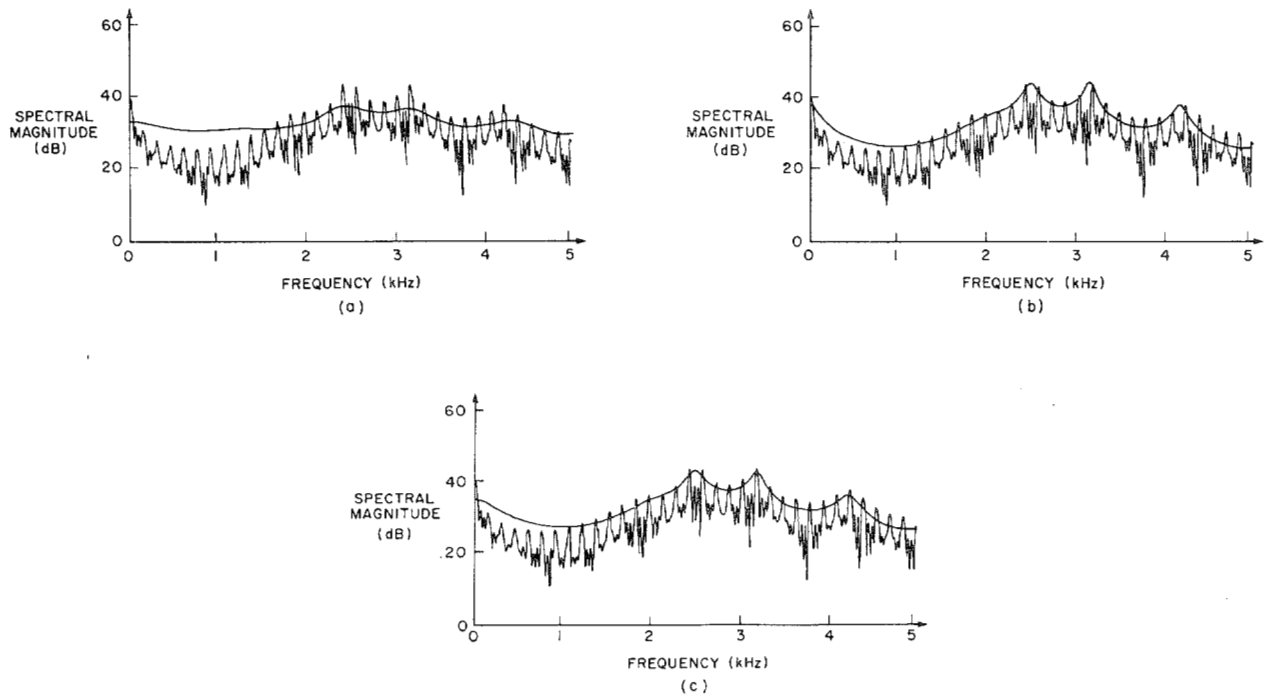

Fig. 7. Same as Fig. 3 with periodic impulse train excitation.

assumption that $s(n) = y(n)$) for (a), by applying System B to noisy synthetic data for (b), and by applying System C to noisy synthetic data for (c). The results shown in (b) and (c) of Figs. 2-4 were obtained after two and ten interactions respectively. In each of the three figures the true spectrum corresponding to the excitation of white noise was included to facilitate the comparisons. Figs. 5-8 are similar to Figs. 1-4, but with a different excitation. Here the excitation is a periodic train of impulses that corresponds to a fundamental frequency of 150 Hz which is typical for an adult male speaker.

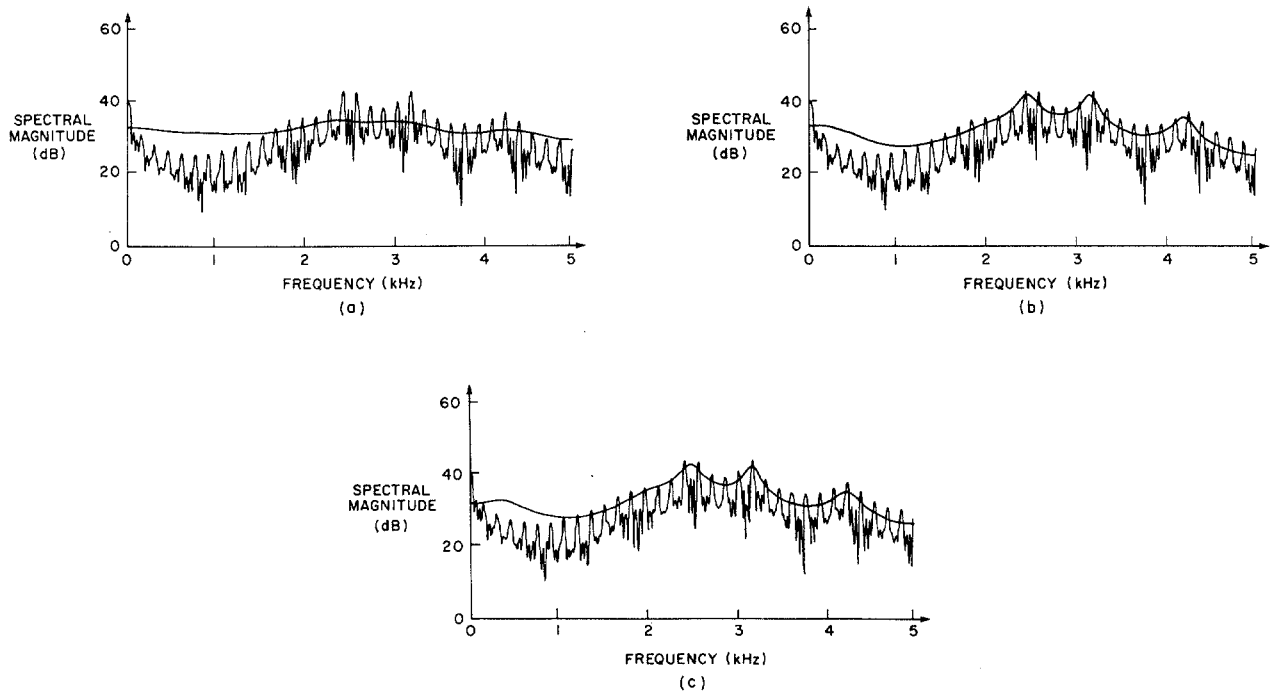

Fig. 8. Same as Fig. 4 with periodic impulse train excitation.

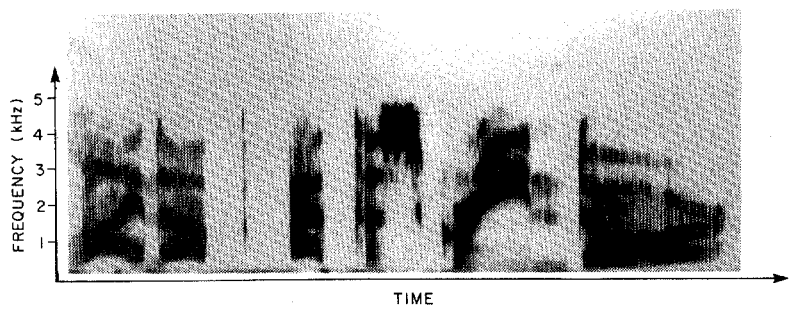

Fig. 9. A spectrogram of an English sentence, "Line up at the screen door," spoken by an adult male speaker.

Again, in (a), (b), and (c) of Figs. 6-8, the true spectrum corresponding to excitation of a periodic train of impulses is included to facilitate the comparisons.

Figs. 9-12 show the results of the analysis based on the real speech data at the S/N ratio of 10 dB. In Fig. 9, a spectrogram of the real speech data with no background noise is shown. Figs. 10-12 correspond to synthesized speech based on the analysis results obtained by applying System A to noisy speech data [with assumption that $s(n) = y(n)$] for Fig. 10, by applying System B with two iterations to noisy speech data for Fig. 11, and by applying System C with ten iterations to noisy speech data for Fig. 12. In all the examples shown (Figs. 10-12), a tenth-order all-pole system was used with the same pitch information (voicing/unvoicing, and pitch period in the case of voicing) obtained from the speech data with no background noise.

Even though our discussions in this section are of a very preliminary nature, we note the following points based on the results of our application of Systems A, B, and C to various other synthetic and real speech data in addition to the results illustrated in Figs. 1-12. First, even though all the theoretical results that lead to System B and System C were based on a stochastic excitation, both systems appear to be applicable, with similar performances to the case when the excitation is a periodic train of impulses. This is consistent with previous results (19) in the clean speech case, where all the theoretical results are obtained based on a stochastic excitation, but experience has shown that the same results can be equally well applied to the case when the excitation is by a periodic train of impulses. Second, in the derivation of System B and System C, we have assumed that s_I is known. When the analysis length N is much greater than

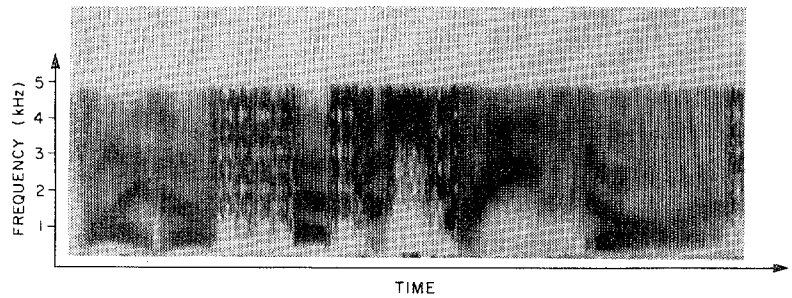

Fig. 10. A spectrogram of the synthesized speech by applying System A to noisy speech data at S/N of 10 dB.

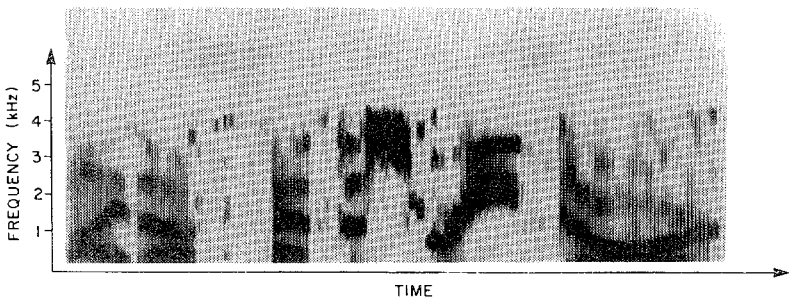

Fig. 11. A spectrogram of the synthesized speech by applying System B to noisy speech data at S/N of 10 dB.

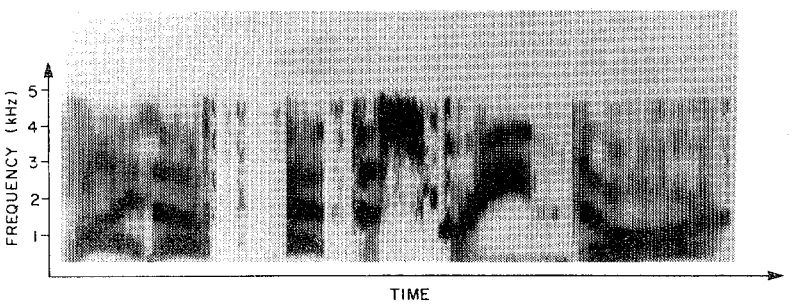

Fig. 12. A spectrogram of the synthesized speech by applying System C to noisy speech data at S/N of 10 dB.

p, the effect of small error in initial conditions is negligible. In fact, the synthetic data used in the above examples were generated with $s_I = \mathbf{0}$ while the use of (41b) in System B and System C is based on the implicit assumption that s_I and s_O are generated by the same stationary process. Results indicate that there is little difference in the performance when slightly different assumptions on s_I are used if N is sufficiently large relative to p. Third, in applying System B to noisy synthetic and real speech data we have shown the results that were obtained after two iterations. We have found that, in general, the converging solution after many iterations generates the vocal-tract transfer function for which the bandwidths of the poles are smaller than those associated with real speech and thus, in the acutal implementation, results obtained after one or two iterations seem to be more desirable. Fourth, in applying System C to noisy synthetic and real speech data, we have shown the results that were obtained after ten iterations. Even though it is not theoretically known whether or not System C has a converging solution, in all our simulations we have empirically observed that System C converges and the results obtained after ten iterations are close to the converging solution. Unlike System B, the converging solution of System C generates the vocal-tract transfer function for which the bandwidths of poles are comparable to those of real speech.

When a sufficiently large amount of noise is added to synthetic or real speech data such that the resulting noisy data have no spectral peaks or spectral peaks that are different

from the pole locations of the original data, the application of System B or System C sometimes generates the vocal-tract transfer function whose pole frequencies are different from those of the original data. This causes some unnatural shift of formant frequencies[3] which can be observed to some extent in Figs. 11 and 12. Unless wrong formant frequencies are estimated sufficiently often, which can happen at very low S/N ratios, we have observed that the primary perceptual effect is generation of "musical tone" like sounds in the background which causes some degradation of speech quality. Within the context of this paper, this problem may be partially solved by introducing some *a priori* information of the all-pole coefficient vector from the past analysis frames. This aspect of improving System B and System C is currently under investigation.

Even though our discussions in this paper are based on white background noise, all the theoretical results in this paper can easily be extended to the background noise with other known spectra. The result of such an extension involves simply substituting $p_d(\omega)$ for σ_d^2 in (41a) and (45) where $P_d(\omega)$ is the power spectrum of the background noise.

Even though the performance of System B and System C can only be properly evaluated by formal subjective tests, our very preliminary informal listening indicates that the two systems are capable of significant noise reduction. A more formal subjective test, which is directed towards evaluating the two systems discussed above in terms of their performance in enhancing speech intelligibility and quality when the background noise is of various different spectra, is currently underway. The results of these tests will be reported in a later paper.

Appendix I

In this Appendix, we briefly summarize the notation that is used in the paper.

$s(n)$: Speech waveform.

$d(n)$: Additive disturbance or background noise, assumed to be zero-mean white Gaussian noise with variance of σ_d^2.

$y(n)$: Noisy speech waveform, $s(n) + d(n)$.

$s(n_1, n_2)$: $\begin{bmatrix} s(n_1) \\ \vdots \\ s(n_2) \end{bmatrix}$.

s_I: $s(-1, -p) = \begin{bmatrix} s(-1) \\ s(-2) \\ \vdots \\ s(-p) \end{bmatrix}$, initial condition vector.

s_O: $s(N-1, 0) = \begin{bmatrix} s(N-1) \\ s(N-2) \\ \vdots \\ s(0) \end{bmatrix}$, speech waveform vector.

y_O: $y(N-1, 0) = \begin{bmatrix} y(N-1) \\ y(N-2) \\ \vdots \\ y(0) \end{bmatrix}$, noisy speech waveform vector.

g: Gain of an all-pole system.

$w(n)$: Zero-mean white Gaussian noise with unity variance.

$\boldsymbol{a}$: $\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$, an all-pole system parameter vector.

$\boldsymbol{a}^T$: $[a_1, a_2, \cdots, a_p]$, transpose of $\boldsymbol{a}$.

$\bar{\boldsymbol{a}}$: *a priori* mean of $\boldsymbol{a}$.

P_0: *a priori* covariance of $\boldsymbol{a}$.

$p(A)$: Probability density function of A.

$p(A \mid B)$: Probability density function of A conditioned on B.

$p(A \mid B; C)$: Probability density function of A conditioned on B and C where C is assumed to be known.

$p(A; B)$: Joint probability density function of A and B where B is assumed to be known.

$p(A)\mid_{A=\hat{A}}$: Probability density function of A evaluated at $A = \hat{A}$. $p(A)\mid_{A=\hat{A}}$ is also denoted by $p(\hat{A})$ when the meaning is clear from the context.

[3]The effect is generally higher at higher formants. This is due to the fact that the effective S/N ratio at higher formants is, in general, lower since speech (voiced sounds) has less energy in the higher frequencies while white noise, which is used as additive background noise in this work, has equal energy at all frequencies.

Acknowledgment

We would like to express our appreciation for a number of helpful discussions to Dr. E. Hofstetter, Prof. N. Sandell, and Prof. Alan Willsky.

References

[1] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 488–494, Dec. 1976.

[2] B. Gold, "Robust speech processing," M.I.T. Lincoln Lab., Tech. Note 1976-6, DDC AD-A012P99/0, Jan. 27, 1976.

[3] V. C. Shields, Jr., "Separation of added speech signals by digital comb filtering," S.M. thesis, Dep. Elec. Eng., M.I.T., 1970.

[4] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. 1976 IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Apr. 12–14, 1976, pp. 251–253.

[5] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Study and development of the INTEL technique for improving speech intelligibility," Nicolet Scientific Corp., Final Rep. NSC-FR/4023, Dec. 1974.

[6] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim, "Evaluation of a speech enhancement system," in *Proc.*

1977 IEEE Int. Conf. on Acoust., Speech, and Signal Processing, May 9-11, 1977, pp. 212-215.
[7] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal Processing*, to be published.
[8] J. S. Lim, "Evaluation of autocorrelation subtraction method for enhancing speech degraded by additive white noise," submitted to *IEEE Trans. Acoust., Speech, Signal Processing*.
[9] P. Eykhoff, *System Identification: Parameter and State Estimation*. New York: Wiley, 1974.
[10] F. C. Schweppe, *Uncertain Dynamic Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1973.
[11] H. L. Van Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968.
[12] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 55, pp. 637-655, 1974.
[13] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561-580, Apr. 1975.
[14] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Heidelberg, New York: Springer-Verlag, 1976.
[15] F. Itakura and B. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," presented at the 6th Int. Cong. Acoust., Y. Kohasi, Ed., Tokyo, Japan, August 21-28, 1968, paper C-5-5.
[16] J. D. Markel and A. H. Gray, Jr., "A linear prediction vocoder simulation based upon the autocorrelation method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 124-134, Apr. 1974.
[17] N. Levinson, "The wiener RMS (root mean square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261-278, 1947.
[18] J. Durbin, "The fitting of time-series models," *Rev. Inst. Int. Statist.*, vol. 28, pp. 233-243, 1960.
[19] A. Gelb *et al.*, *Applied Optimal Estimation*, A. Gelb, Ed. Cambridge, MA: M.I.T. Press, 1974.
[20] J. D. Gibson, J. L. Melsa, and S. K. Jones, "Digital speech analysis using sequential estimation techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 362-369, Aug. 1975.
[21] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1965.
[22] K. Steiglitz and L. E. McBride, "A technique for the identification of linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-10, pp. 461-464, 1965.
[23] J. K. Åström and P. Eykhoff, "System identification--A survey," *Automatica*, vol. 7, pp. 123-162, 1971.

Quasi-Periodic Instability in a Linear Prediction Analysis of Voiced Speech

D. G. NICHOL, MEMBER, IEEE, AND R. E. BOGNER, MEMBER, IEEE

Abstract—**A significant semiperiodic fluctuation of the vocal tract area functions derived by linear prediction of the speech waveform has been noted during apparently stationary voiced segments of speech. In one example some values of the area function varied over a range of 9 : 1 over a few pitch periods. The phenomenon is attributed to "beating" of the pitch period and the time interval between successive computations which causes variations of the time relationship between glottal pulse and analysis window. This is supported by the fact that no fluctuations occur in the area function derived from natural or synthetic speech when the computation interval is equal to the pitch period. Any slight difference between the two leads to significant pulsations, however. A simple theoretical model is used to show how the positioning of the analysis window can influence area function estimates.**

The problem can be largely overcome by using longer time windows (greater than 2.5 pitch periods), or alternatively, by averaging the area functions over several adjacent intervals.

Manuscript received August 1, 1977; revised December 16, 1977, and January 18, 1978. This work was supported in part by the Department of Defence and the Australian Research Grants Committee.

D. G. Nichol is with the Weapons Research Establishment, Adelaide, South Australia.

R. E. Bogner is with the Department of Electrical Engineering, University of Adelaide, Adelaide, South Australia.

I. INTRODUCTION

SEVERAL attempts have been made recently to use linear prediction analysis of speech for isolated-word [1], [2] and spoken-digit recognition [3], [4]. The feature chosen for the recognition algorithm in these studies was the set of linear prediction coefficients. It is well known that an estimate of the vocal tract area function can be derived from these coefficients [5]-[7] and the present paper arose from a study of the usefulness of this function for both speech and voice recognition. Because of the extensive information available from phonetic and articulatory studies of speech production, it was believed that the vocal tract area function (VTAF)

0096-3518/78/0600-0210$00.75